Clinical AI tools must convey predictive uncertainty for each individual patient

Christopher R. S. Banerji[1,2], Tapabrata Chakraborti[1,3], Chris Harbron[4], Ben D. MacArthur[1,5,6]

[1] The Alan Turing Institute, London, UK

[2] University College London Hospitals, NHS Foundation Trust, London, UK.

[3] UCL Cancer Institute, Faculty of Medical Sciences, University College London, London, UK.

[4] Roche Pharmaceuticals, Welwyn Garden City, UK

[5] Faculty of Medicine, University of Southampton, Southampton, UK

[6] Mathematical Sciences, University of Southampton, Southampton, UK

Stand first

AI tools usually aim to maximise predictive accuracy, but personalised measures of uncertainty, using new techniques such as conformal prediction, are needed for clinical AI to realise its potential.

Main text

Artificial intelligence (AI) is a powerful and rapidly developing technology with the potential to revolutionise personalised medicine and dramatically improve human health[1]. Yet, although thousands of papers are published each year on AI in healthcare, it remains strikingly absent from clinical practice – a fact that is widely acknowledged, but incompletely understood[2]. Unlocking the clinical potential of AI requires a firmer understanding of the strengths and weaknesses of AI tools in the clinic, and the technical, clinical, economic, and sociological barriers to their use in healthcare contexts.

Population-level accuracy

There are many reasons for lack of translation of AI to the clinic, including biased study design[3], poor clinical performance[2], clinician scepticism[4], and concerns over patient data misuse[5]. A more fundamental issue is that the standard pipelines used to develop AI tools are often adapted from other disciplines, such as computer vision and natural language processing, and are designed to optimise population level accuracy metrics that are meaningful for those disciplines but are mismatched to the clinical context. AI tools built in this way do not capture the vital clinical fact that each patient is a unique individual.

Patient populations are diverse. Individual differences in drug metabolism can give rise to serious side-effects in some patients but be lifesaving in others[6]; women with myocardial infarction have a different clinical presentation to men[7]; prognosis in most malignancies is linked to personalised characteristics[8].

AI models anchored on population-level accuracy metrics risk undervaluing these differences. While they may be accurate on average, such tools cannot provide reliable advice for all individual patients. When a patient, enquiring about their own health, asks their clinician ‘Are you sure?’ such AI tools cannot consistently provide an accurate answer.

The ability of clinicians to convey uncertainties to patients with sensitivity is central to clinical medicine. Taking AI tools into the clinic will therefore require new approaches to model development,

that use clinically relevant metrics[2] , recognise the distinctiveness of each individual patient, and provide personalised measures of performance. Successful clinical AI tools cannot simply maximise predictive accuracy; they must also convey uncertainty.

When assessing an AI model for accuracy, for example, to diagnose a new patient or offer a prognosis, we measure how close its predictions are to the truth, typically using a characterised patient test population for whom clinical outcomes are known. However, when applying AI in the clinic, the truth is *a priori* unknown and so the clinician must also assess the factors that interfere with the model's ability to make an informed decision. Understanding these uncertainties, and how the model has handled them, is much more useful to the clinician and meaningful to the patient than providing a single "most likely" recommendation.

Personalised uncertainty

In the data science literature, two types of uncertainty are commonly considered[9]: epistemic (or knowledge) uncertainty arises when clinical knowledge (or an AI model) does not fully capture the relationships between the patient's data and the clinical outcome being predicted. Since knowledge and models can be updated with new evidence, epistemic uncertainty can potentially be reduced, for instance, by carrying out additional diagnostic tests. By contrast, aleatoric (or data) uncertainty arises from measurement variations due to the inherent uniqueness of every individual, as well as the randomness or noise associated with the data capture processes. Aleatoric uncertainty therefore cannot be fully eliminated.

Although useful, this dichotomy is somewhat artificial, since epistemic and aleatoric uncertainties may be (and often are) interrelated. For example, as technology improves, new and varied ways to collect ever-more healthcare data are developed, from spatial transcriptomics to wearable devices. While these new sources of information can provide knowledge about healthcare outcomes and reduce epistemic uncertainty, they may also be imprecise, and this imprecision can increase aleatoric uncertainty. As patients are empowered to monitor their own health, and inevitably do so inconsistently, such data-driven uncertainty may vary depending on the patient and become increasingly personalised.

The ability to distinguish between different kinds of uncertainty is important in healthcare because clinical decision-making requires knowledge of not only how uncertain a model is, but also why it is uncertain. By revealing data or model weaknesses, properly quantified uncertainty can therefore be profoundly beneficial, and provide rationale for model improvement over time.

Uncertainty drives improvements

Many poignant examples of uncertainty driving improved clinical care arose during the COVID-19 pandemic, when the emergence of a new, uncharacterised pathogen introduced significant epistemic uncertainty. For example, a shift in associations between clinical variables led to spurious alarms in automated sepsis alert systems. These false positives were so ubiquitous that some hospitals decommissioned their alert systems[10]. A subsequent meta-analysis showed that as many as 75% of COVID-19 patients were prescribed antibiotics despite only 8.6% having bacterial co-infection[11]. The reduced ability to discriminate serious bacterial infection from SARS-CoV-2 infection drove widespread investigation of biomarkers for antimicrobial stewardship[12], improving clinical knowledge, reducing epistemic uncertainty and positively impacting patient care.

Similarly, in a non-clinical setting, it has been long recognized that a range of facial recognition algorithms can be highly accurate in identifying lighter-skinned, male faces but are uncertain in recognising darker-skinned, female faces[13].This uncertainty is due to under-representation of darker-skinned subjects in training data, which prevents AI models from accurately learning features related to this group. Because this is an issue of epistemic uncertainty, it can be addressed by more scrupulous collection of representative data, and careful training of models[13,14]. In this case model

uncertainty again imparts useful information since it highlights shortcomings in the data collection processes that can, and should, be rectified.

Conformal prediction

Good clinicians see uncertainty as an opportunity to deepen their understanding and improve care of their patients. For AI tools to be truly useful in the clinic, they must be used judiciously by clinicians, to identify and quantify sources of uncertainty, understand how they affect clinical outcomes and thereby improve clinical decision-making.

The idea that there is information in uncertainty is an emerging and rapidly growing theme in modern data science[15] and in recent years new tools have emerged that can produce personalised measures of uncertainty[16–19]. One suite of methods, known as conformal prediction, may provide the bridge needed to take clinical AI from theory to practice.

Conformal prediction encompasses a set of robust statistical processes to convert a heuristic notion of uncertainty, determined at the population level, to a rigorous assessment of uncertainty for each individual. To illustrate how conformal prediction works, consider diagnosing the cause of a headache from a set of clinical variables (**Figure 1**). The standard AI approach to this problem is to build a computational model using a training dataset to learn how patterns in clinical variables associate with different underlying pathologies. Using this trained model, any new patient presenting with a headache can be offered a diagnosis depending on how the trained model interprets their clinical presentation.

When making judgements, AI models typically produce measures of their certainty. For instance, our AI model may be 95% sure that the underlying cause of headache for a given patient is migraine. However, these measures of uncertainty are only heuristic: they do not concern the true diagnosis of the individual patient, but rather the model's own internal assessment of the patient's characteristics, based on the data it has seen previously. While apparently useful, these measurements of certainty can therefore be misleading. For example, if the model is poorly trained it may offer confident, but incorrect, advice. For complex "black box" models that are not easily interpreted, the reasons for this confidence may also be opaque. Conformal prediction is designed to resolve this issue.

Confidence through experience

The inventors of conformal prediction, Alex Gammerman and Vladimir Vovk, have said that "conformal prediction uses past experience to determine precise levels of confidence in new predictions". This has a clear parallel to learning through clinical practice. The technique compares predictions made by an AI model for a new observation to the predictions of a group of similar observations for whom outcomes are known. It does so by calibrating the model's predictive accuracy against its level of certainty for all possible outcomes. This process translates the output of the model from a single most-likely outcome to a shortlist of possible outcomes that are all feasible. Using the example of a patient presenting with headache, rather than being assigned one diagnosis (e.g., the patient most likely has migraine), using conformal prediction each new patient presenting with a headache is provided a list of personalised possible diagnoses, depending on the model's heuristic certainty for each possible outcome.

Subject to some general assumptions, this shortlist of possible diagnoses is guaranteed to contain the patient's true diagnosis with a level of confidence that can be specified by the clinician that reflects the clinical situation (e.g., 95%) – regardless of the properties of the data or kind of AI model used to make predictions[20]. Conformal prediction calibrates model accuracy against certainty, and so if the model is either unsure of the correct clinical diagnosis or is over-confident, then the list of possible diagnoses will be long, which indicates to the clinician that more information is needed to correctly identify the underlying pathology and appropriately treat the patient. Conversely, if the model is confident, and this confidence is based on reliable clinical evidence, then the list will be short, giving the clinician confidence of an accurate diagnosis (**Figure 1**).

When making decisions, clinicians must consider not only how likely a diagnosis is, but also how consequential it would be for the patient. Conformal prediction can also be adapted to account for severity of outcome, so that relevant serious conditions are “upweighted” and appropriately presented to the clinician, even if they are less likely. By so-doing, conformal prediction shifts the focus of the AI model from trying to find one accurate clinical recommendation to offering the clinician a range of possibilities, tailored to the individual patient, which can then be investigated further. Not only is this technically more robust, it also circumvents many of the practical hindrances to deployment of AI into the clinic, providing confidence in AI-enabled decisions both for clinicians and for patients.

Widespread use of AI in the clinic will require AI tools to move away from simply maximizing predictive accuracy toward harnessing uncertainty, and supporting the clinician in making well-informed decisions for each individual patient. New methods such as conformal prediction are making this transition possible and will be critical to truly integrating AI into clinical practice, helping doctors improve healthcare for all patients.

References

1. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nature Medicine 2022 28:1* **28**, 31–38 (2022).

2. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* **17**, 1–9 (2019).

3. Andaur Navarro, C. L. *et al.* Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* **375**, 2281 (2021).

4. Gaube, S. *et al.* Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine 2021 4:1* **4**, 1–8 (2021).

5. Murdoch, B. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Med Ethics* **22**, 1–5 (2021).

6. Evans, W. E. & Relling, M. v. Pharmacogenomics: Translating functional genomics into rational therapeutics. *Science (1979)* **286**, 487–491 (1999).

7. Coventry, L. L., Finn, J. & Bremner, A. P. Sex differences in symptom presentation in acute myocardial infarction: A systematic review and meta-analysis. *Heart & Lung* **40**, 477–491 (2011).

8. Jackson, S. E. & Chester, J. D. Personalised cancer medicine. *Int J Cancer* **137**, 262–266 (2015).

9. Kiureghian, A. der & Ditlevsen, O. Aleatory or epistemic? Does it matter? *Structural Safety* **31**, 105–112 (2009).

10. Finlayson, S. G. *et al.* The Clinician and Dataset Shift in Artificial Intelligence. *New England Journal of Medicine* **385**, 283–286 (2021).

11. Langford, B. J. *et al.* Antibiotic prescribing in patients with COVID-19: rapid review and meta-analysis. *Clinical Microbiology and Infection* **27**, 520–531 (2021).

12. Heesom, L. *et al.* Procalcitonin as an antibiotic stewardship tool in COVID-19 patients in the intensive care unit. *J Glob Antimicrob Resist* **22**, 782–784 (2020).

13. Buolamwini, J. & Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. in *Proceedings of Machine Learning Research* vol. 81 77–91 (PMLR, 2018).

14. Merler, M., Ratha, N., Feris, R. & Smith, J. R. Diversity in Faces. (2019).

15. Mitra, R. *et al.* Learning from data with structured missingness. *Nature Machine Intelligence 2023 5:1* **5**, 13–23 (2023).

16. Vovk, V., Hoi, S. C. H. & Buntine, W. Conditional validity of inductive conformal predictors. in *Proceedings of the Asian Conference on Machine Learning 25* vol. 25 475–490 (2012).

17. Barber, R. F., Candes, E. J., Ramdas, A. & Tibshirani, R. J. Conformal prediction beyond exchangeability. Preprint at http://arxiv.org/abs/2202.13415 (2022).

18. Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L. & Schuster, T. Conformal Risk Control. Preprint at http://arxiv.org/abs/2208.02814 (2022).

19. Tibshirani, R. J., Barber, R. F., Candes, E. J. & Ramdas, A. Conformal Prediction Under Covariate Shift. Preprint at http://arxiv.org/abs/1904.06019 (2019).

20. Shafer, G. & Vovk, V. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research* **9**, 371–421 (2008).